# 感知不透明性增加职场中的算法厌恶*

赵一骏 许丽颖 喻丰 金旺龙

（武汉大学心理学系，武汉 430072）

**摘　要**　职场中用算法作为人类决策的辅助和替代屡见不鲜，但人们表现出算法厌恶。本研究通过 4 个递进实验在不同职场应用场景下比较了人们对于人类决策者与算法决策者所做决策的态度，并探讨其内在机制和边界条件。结果发现：在职场情境中，相比于人类决策者，人们对算法决策的可容许性、喜爱程度、利用意愿更低，表现出“算法厌恶”。这一现象的内在心理机制是相比于人类决策，人们认为算法决策者的决策更加不透明（实验 2~3）。进一步研究发现，当算法被赋予拟人化特征时人们扭转了对算法决策的厌恶，提高了对其的接纳态度（实验 4）。研究结果有助于更好地理解人们对算法决策的反应，为推动社会治理智能化、引导算法使用伦理化提供启示。

**关键词**　算法厌恶，透明性，拟人化，职场

**分类号**　B849: C91

## 1 引言

2022 年 8 月 Facebook 母公司 Meta 使用算法裁掉了 60 名劳务派遣员工，不少员工表示，除了听说自己是被算法“随机”解雇之外，并未收到其他的任何解释（Hays, 2022）。无独有偶，Meta 并非第一个使用算法进行人力资源管理的企业。早在 2015 年，Amazon 针对仓储管理开发了一套人工智能效率检测和评估系统，用来实时监控员工的工作状态，并将其纳入绩效考核之中（Upadhye, 2018）。2021 年 8 月，俄罗斯一家在线支付服务公司 Xsolla 同样使用算法评估出的“数字足迹（digital footprint）”解雇了 150 名存在态度问题和效率低下的在职员工（McAloon, 2021）。职场中的算法使用问题一时间成为社会热点。近些年，越来越多的企业将算法应用于职场环境之中，这提高了组织效率，承担并解放了原先由人力资源管理者来完成的繁杂工作和职责（Basu et al., 2023; Brynjolfsson & McAfee, 2014; Chalfin et al., 2016; Duggan et al., 2020; Garg et al., 2022）。据在线招聘平台 Career Builder 的调查，美国约 55%的人力资源经理认为以算法为核心的人工智能系统会在未来几年内成为他们工

收稿日期：2023-12-20

* 国家自然科学基金青年项目(72101132); 国家社科基金青年项目(20CZX059)

通信作者：许丽颖, E-mali: liyingxu830@gmail.com; 喻丰, E-mail: psychpedia@whu.edu.cn

作的常规部分（HR Daily Advisor Staff, 2017）。由此可见，以算法为代表的智能决策体系融入职场逐渐从构想成为现实，并大为发展。

与算法进入医疗诊断（Hao, 2020）、金融投资（Ahmed et al., 2022）、司法量刑（Hao, 2019）、交通驾驶（Badue et al., 2021）等社会生活领域的原因一致，算法之所以能够迅速席卷职场，主要是因为传统的人类决策有其自身无法突破的局限。在传统模式中，作为决策主体的人，极易受到先前经验（Kahneman & Tversky, 1972）、情绪（Lerner et al., 2015）等干扰因素的影响而出现决策失误，造成不良后果。然而随着大数据、深度学习、神经网络等计算机技术的蓬勃发展，算法成为了新的决策主体。算法在进行决策的时候，具有对海量信息强大的统摄能力（Blair & Saffidine, 2019），且决策速度快（Bonnefon et al., 2016）、客观公正（Andrews et al., 2006）、并且不会受到疲劳和情绪等因素的干扰（Barnes et al., 2015）。也正因此，算法被广泛地应用于提供建议、判断和预测领域，形成算法决策（algorithmic decision making）的概念（Burton et al., 2020）。本研究主要关注人们对职场中算法使用的态度，并进一步探索其中可能的内在原因和边界条件。

### 1.1 职场中的算法厌恶

尽管算法决策表现出上述种种优势，为社会效率带来飞跃式提升，但人们似乎并没有因为其强大的算力和准确的预测能力而轻易地接纳算法，反而对这一方兴未艾的新兴技术持有一种普遍的悲观态度（Jussupow et al.,2020; Liu et al., 2023; Mahmud et al., 2022）。部分企业家从经营实务的经验出发，认为自动化崛起将是人类“最大的生存威胁”（McFarland, 2014）。如特斯拉创始人 Elon Musk 提出，人工智能正在召唤“恶魔”的出场（Lemieux, 2017）。前谷歌首席执行官 Eric Schmidt 表示担心人工智能会给人类带来一种“生存风险”（Kharpal, 2023）。哲学家们也对算法决策的兴起存在着一种形而上式的担忧，认为机器代替人类做决定可能会导致一场彻底的灾难（Bostrom, 2014）。这场灾难可能是存在论意义上的危机，即一旦人工智能诞生出自我意识，人无异于创造出了一个能够亲手毁灭人类历史、否定人类存在价值的技术“上帝”（赵汀阳, 2018; 2019）。对于大众来说，尽管算法通常能够比人类更准确地完成决策任务，人们依然会倾向于选择接纳人类的决策而不是算法提出的决策，这种倾向性也被称之为“算法厌恶（algorithm aversion）”（Dietvorst et al., 2015）。这种厌恶倾向实质上是对算法的一种认知偏见，并全面地表现在“知、情、行”等方面（张语嫣 等, 2022），即在认知上持有否定态度（e.g., Prahl & van Swol, 2017）、在情感上表露嫌恶情绪（e.g., Lee, 2018）、在行为上存在拒绝倾向（e.g., Filiz et al., 2021）。算法厌恶同时也表现在绘画艺术判断（Millet et al., 2023）、诗歌美学判断（Hitsuwari et al., 2023）、医疗决策（Longoni et al.,

2019）、战争和司法决策（Bigman & Gray, 2018）、日常电影和书籍推荐（Longoni & Cian, 2022; Yeomans et al., 2019）等领域。

在职场中，算法被认为可以在一定程度上消除无意识的人类偏见（Cheng & Hackett, 2021），从而可能做出更加平等、不被干扰的决策。已有研究发现，在普通工作体量下，算法决策在 80%以上的情境中是优于人类管理者决策的（Yu et al., 2017）。但即便有此优势，人们依然倾向于认为算法管理是一种“暴政”，就如同“时钟暴政”一般压榨、剥削着员工（Lehdonvirta, 2018）。被算法日趋自动化的工作会剥夺人们对自身工作的控制感（Holford, 2022），从而降低人在其中所收获的自主性和责任感（Goods et al., 2019; Langer et al., 2021）。并且，算法管理还会通过人们在工作中体验到的自主性和对工作所获种种奖励的期望降低幸福感（Kinowska & Sienkiewicz, 2022）。就晋升、解雇、分配年终奖等管理实践中的具体场景而言，人们会倾向于认为算法主体的决策仅仅是量化的、去背景化的，而无法考虑到质性和环境因素并由此做出周全的决策，所以会认为算法决策（与同样的人类决策相比）是不公平的，甚至因此降低了对组织的情感承诺（Newman et al., 2020）。就招聘而言，相比于算法，应聘者更信任由人类管理者来进行简历筛选或面试选拔，这是因为人类管理者的筛选被认为是更可控、公平的（Langer et al., 2019）。而这一现象在应聘者承诺契约时同样有所表现，即相比于招聘人员向应聘者许诺丰厚的薪资、奖金，当他们想通过向你保证你会从工作中收获成长、感受融洽氛围达成合约时，人们更拒绝由算法提出（Tomprou & Lee, 2022）。同时，员工会对应用自动化算法的人力资源管理感到包括情感、心理、隐私、社交在内的六种负担，可一旦人类决策者能够参与并把关的话，这种负担就会一定程度地减轻（Park et al., 2021）。

因此，本研究提出假设 H1：相比于人类人力资源管理者（下文均简称为 HR），人们倾向于在认知上拒绝、在情感上厌恶、在行为上回避算法 HR 在职场中做出有关决策。

### 1.2 感知透明性

面对不同主体的决策，究竟是什么因素造成人们对其产生不同的反应？透明性（transparency）是其重要原因。算法透明性意味着对于用户来说可以理解算法系统正在做什么，以及为何这样做（IEEE, 2017）。算法透明性所强调的是对于一般使用者而言，智能系统的决策过程与机制要具有可知性和可解释性，决策后果要具有可预测性和可说明性，即决策过程对使用者而言是可以轻易理解的（Nefdt, 2020）。透明性之所以如此重要，是因为人们对于决策结果的反应在一定程度上取决于决策者所提供的信息以及其能否被解释（Dodge et al., 2019），并且人们希望决策者提供他们能够理解的决策（McNee et al., 2006;

Herlocker et al., 2004)。而事实上，算法系统对人的认知而言却是一个黑箱（black box）（Burrell, 2016; Nicholson Price, 2018），即算法被人们知觉为一个工作机制神秘且复杂的系统，只能简单地观测到它的输入端和输出端，而无法得知算法究竟是如何处理这一过程并得到结论的（Pasquale, 2016）。形成算法黑箱的原因可能是运算过程本身不可预测，亦可能是人们缺乏相关知识经验（Kroll et al., 2017）。

但不透明的黑箱性质成为了人们接纳算法的阻碍之一。由于算法缺乏自主解释其决策理由的能力，人们可能会对其产生不公平的理解和不安的感受（Acikgoz et al., 2020; Langer et al., 2019）。并且，研究发现算法透明性能够正向预测算法满意度，也就是说，人们认为算法的决策过程越透明，就会对此算法提供的服务越满意（Shin & Park, 2019; Shin, 2020）。此外，倘若使用户更了解系统决策的工作原理，人们便会对其拥有更多的信任（Lee & Boynton, 2017）。这说明当“算法黑箱”被打开后，人们可能会改变对其固有的负面反应。同时，人们对人类决策的理解通常会展现出过度自信（Chen et al., 2023; Moore & Healy, 2008），即人们过高地估计了自己对于他人心理过程的理解程度。这种偏差的产生可能源于人们直觉内省的将心比心过程（Nisbett & Wilson, 1977），可实际上人们在理解他人决策时更多采用的只是直觉（Dane et al., 2012）和启发式（Kahneman, 2003）。即使在颇具专业知识的医学领域，普通人也自以为很了解同为人类的医生是如何诊断癌症的（Cadario et al., 2021）。

可见，人们对不同的决策主体有不同的感知透明性，而对决策过程的理解又能极大程度上影响人们对于决策的态度。据此，本研究提出假设 H2：感知透明性在职场决策主体（人类 HR vs. 算法 HR）对决策反应的效应中起中介作用。

### 1.3 拟人化

拟人化（anthropomorphism）是将人的属性、特征、意向性、心理状态等赋予非人实体的现象（Epley et al., 2007; 许丽颖 等, 2017）。其中，将一个实体视作具有诸多心智能力的过程是拟人化的重要构成因素之一（Waytz et al., 2010）。思维过程的可理解性乃是诸多心智能力的其中之一，因此当我们认为一个实体的思维活动是可以理解、通达的时候，其实我们就是为它赋予了类人的心智（Gray et al., 2007; Gray et al., 2012）。赋予原本不具可理解性的非人算法以人类独有的感知透明性的过程，自然等同于拟人化的过程。因此透明性构成了人力算法与人力资源管理者之间的重要差别，并使得前者被排除在了具有心智能力的实体范围之外。可以预测的是，倘若将非人算法进行外观（De Visser et al., 2016）、声音（Adam et al., 2021）、动作（Fraune, 2020）、心理能力（Moussawi & Koufaris, 2019）等不同层次的

拟人化操纵，使算法在人的感知中越来越像人，弥合算法与人存在于感知中的差别，或许人们对其的态度会更加积极。先前研究也发现，对机器、算法、人工智能拟人化能够降低人们的厌恶倾向。如对自动驾驶汽车进行表层拟人化（即赋予其类人外观或名称），能够有效提高人们对其的信任（Wu et al., 2023）。再或者相比于声音机械、回应冷漠的人工智能助手，一个具有类人外观、拟人姓名、似人声音、以及有温度的情感反应的 AI 助手，会被使用者认为在心理上有更近的距离，并对其服务给予更高的满意度评价（Li & Sung, 2021）。甚至当机器人犯错做出不好的决断时，如果面对一位具有拟人化特征的机器人，人们会认为其具有更高水平的体验性（experience）能力，会因此降低对其失误的负面评价，并在之后选择原谅（Yam et al., 2021）。于是，拟人化促使人们将算法视为一个有心智的主体，并对其采取更加似人的评价和反应，进而在一定程度上影响算法厌恶。特别是对于具有黑箱性质的算法而言，将其拟人化意味着将本不属于非人实体的人类特征赋予它，这在一定程度上拉近了算法与人类两个决策主体之间的差异，使人们对算法的认识更靠近原本对于人类的认识，从而有可能在考虑算法决策有关事宜时更趋近原先对人类决策的态度，进而可能降低算法厌恶的倾向（许丽颖 等, 2022）。在职场中，算法决策原本是模糊、被排斥、并不被允许做出相关决策的，而如果赋予其一些人类特质，将原本非人的算法改造为更贴近人类的拟人化算法，人们在接受其所给出的决策时就理应会有更积极反应。

于是，本研究提出假设 H3：算法拟人化在职场决策主体（人类 HR vs. 算法 HR）对决策态度的影响中起调节作用。这意味着相比于非拟人化算法，人们更加接纳拟人化算法 HR 的决策。

### 1.4 研究概览

基于以上论述，本研究进行了 4 个系列实验，试图考察人们对职场中算法 HR 决策的态度是否有别于对人类 HR 决策的态度，并在此基础上进一步探讨其心理机制和边界条件。基本假设是：相比于传统的人类 HR，人们对算法 HR 决策的态度更加厌恶，态度具体则指认知上不容许、情感上不喜欢、行为上不愿再使用算法做相似决策。这一效应将会受到感知透明性的中介和拟人化的调节。本研究采用递进的 4 个情境实验来探索此假设的有效性，涉及到的职场决策包括招聘录用（实验 1）、年终奖分配（实验 2）、简历筛选（实验 3）和绩效考核（实验 4），并涵盖了具有代表性的全国范围被试和大学生被试。

具体而言：实验 1 探索主要假设，即人们表现出对职场中算法 HR 决策的厌恶反应。实验 2 探究其中的潜在的心理机制，试图发现感知透明性在决策主体影响决策态度中的中介作用。实验 3 通过操纵算法决策的透明度，进一步考察感知透明性是否是导致人们对算法 HR

决策产生厌恶的前因。实验 4 探索职场中人力决策主体对人们的决策态度影响可能的边界条件，探求拟人化在决策主体影响决策态度中的调节效应。

# 2 实验 1：职场中的算法厌恶

实验 1 的目的是初步探讨与人类HR相比, 人们对算法HR所做决策的态度是否更加厌恶。我们采用情境实验的方法, 将被试随机分配至人类组和算法组, 分别阅读人类 HR 和算法HR做出决策的情境材料并报告对其决策的可容许性、喜爱程度和对该HR的利用意愿, 以此比较人们对人类 HR 和算法 HR 决策的态度差异，验证职场中算法厌恶的存在。

## 2.1 方法

### 2.1.1 被试

本研究首先使用 G*Power 3.1 软件（Faul et al., 2009）计算研究所需样本量。以独立样本 $t$ 检验为统计方式，显著性水平$\alpha = 0.05$ 且中等效应量($d = 0.5$) 时，为了达到 90%统计检验力，本实验至少需要 172 名被试。为了确保最终有足够的数据用于分析，我们通过问卷星平台在某高校内发放了 416 份问卷，排除未通过注意力检测题目的 72 份问卷后得到 344 份问卷，问卷回收率为 82.69%。再排除未通过操纵检查 41 份问卷后，最终得到 303 名被试的问卷用于统计分析（总回收率为 72.84%），其中男性 125 名(41.3%)，女性 178 名(58.7%)，平均年龄 $M = 20.80$ 岁，$SD = 1.61$ 岁。参与实验的被试被随机分派到人类组和算法组，其中人类组 163 人，算法组 140 人。所有被试均自愿参加实验且知情同意，通过注意检查的被试在实验结束之后获得相应实验报酬。

### 2.1.2 实验设计与程序

实验 1 为单因素两水平被试间实验设计，两组分别为人类组和算法组，所有被试随机分配到其中一组，首先阅读对应的人类 HR 或算法 HR 做决策的情境材料。

被试阅读的材料为（决策主体的变化通过字体加粗显示）：“在去年的秋季招聘中，**九日集团的人力资源总监王鹏负责/九日使用“330F”招聘算法对投递**简历的应届毕业生进行评估。整个招聘过程，由**王鹏总监带领他的团队/算法“330F”**独立完成、全权决定、并对招聘结果负责。**王鹏团队是业内资深的人力评估团队/“330F”招聘算法是一种新型招聘算法**，能够仔细审查应聘者的简历和背景，准确预测未来可以满足工作岗位需求的员工、适合企业文化的员工，能够从海量应聘者中找出最适合该企业该岗位的员工。据了解，参与面试的应届毕业生有 50 人，但最终经过**王鹏团队/“330F”招聘算法**的决定，仅录取 5 人（满足九日公司要求的 10%通过率），名单如下：王圣义、黄焦旭、陈振江、谢文祥、许翰芸，

**王鹏团队/“330F”招聘算法**第一时间将录用名单上传并公示在招聘网站上，但并未对此进行其他说明（包括评估分数排名、评估细则等等）”。

以上材料改编自 Newman 等人（2020）的研究，两组被试所阅读的材料仅更换了决策主体的身份（即进行招聘筛选并做出最终决定的是王鹏领导的人类资源团队或招聘算法“330F”），而不作其他任何更改。为了确保被试认真阅读并理解了情境材料的内容, 被试在阅读完情境材料后被要求回答注意力检查题目，即人类组被试回答“九日集团实施招聘的决策主体（决策主体指：参与、主导、执行决策的实体，是决策系统的灵魂和核心）是资深的人力资源总监——王鹏及其团队吗？”，算法组回答“九日集团实施招聘的决策主体（决策主体指：参与、主导、执行决策的实体，是决策系统的灵魂和核心）是新型招聘算法“330F”吗？”，被试可选择回答“是”或“否”，回答“否”的被试则视为未能通过注意力检查。

在阅读完情境材料并进行注意力检查后，被试被要求填写了对上述决策可容许性、喜爱程度和利用意愿的有关测量问题。其中可容许性（permissibility）的测量改编自 Bigman 和 Gray（2018）的有关测量，要求被试回答以下 3 个题目（括号中为算法组题目）：“王鹏（算法“330F”）做出的决定是否合适？”（1~5 评分，1 代表完全不合适，5 代表完全合适）；“王鹏（算法“330F”）是否应被允许做出这些决策？”（1~5 评分，1 代表完全不应被允许，5 代表完全应被允许）；“王鹏（算法“330F”）是否应被禁止做出这些决策？”（1~5 评分，反向计分，1 代表完全不应被禁止，5 代表完全应被禁止）。3 个题目采用李克特 5 点量表计分，第三个题目为反向计分题目，三题总得分越高表明被试对情境中人类（算法）决策的可容许性越高。实验 1 中该可容许性测量的内部一致性信度 Cronbach’s $\alpha = 0.76$。

被试还被要求回答两题用于测量其对决策主体所做决策的喜爱程度的题目，该两题改编自 Jago（2019）的研究，采用李克特 7 点量表计分，分别为：“你有多赞成王鹏（算法“330F”）所做出的上述决策？”（1~7 计分 1 代表完全不赞成，7 代表完全赞成）；“你有多喜欢王鹏（算法“330F”）所做出的上述决策？”（1~7 计分，1 代表完全不喜欢，7 代表完全喜欢）。2 个题目采用李克特 7 点量表计分，得分越高表明被试对情境中人类（算法）所做的决策越喜爱。实验 1 中两题目间相关系数 $r = 0.701$，$p < 0.001$。

此外，被试还被要求报告了其对于该决策主体的利用意愿，即“如果你是一位企业负责人，你在多大程度上会聘用王鹏及其团队（智能算法 330F）来完成上述决策工作？”，该题目改编自 Cadario 等人（2021）的相似测量，并且同样采用李克特 7 点量表计分（1~7 计分，从“1 = 完全不可能”到“7 = 极有可能”），得分越高表明被试对情境中决策主体的利用意愿更强。

最后，被试报告了性别和年龄两项人口统计学信息。另外，在问卷填写过程中还有两道随机出现的注意力检查题目（如：此题请选择 1）用于筛选未认真作答的被试。

## 2.2 结果

独立样本 $t$ 检验结果表明，人类组被试对决策主体所作决策的可容许性评分（$M = 10.60$，$SD = 1.84$）显著高于算法组（$M = 9.25$，$SD = 2.38$），$t(301) = 5.56$，$p < 0.001$，Cohen's $d = 0.63$。同时，人类组被试对决策主体所作决策的喜欢程度评分（$M = 7.89$，$SD = 2.10$）也显著高于算法组（$M = 7.3$，$SD = 2.10$），$t(301) = 2.08$，$p = 0.038$，Cohen's $d = 0.24$。此外，以利用意愿作为因变量，发现人类组被试对人类 HR 的利用意愿（$M = 4.04$，$SD = 1.42$）高于算法组被试对算法 HR 的利用意愿（$M = 3.74$，$SD = 1.50$），$t(301) = 1.75$，$p = 0.081$，Cohen's $d = 0.21$，差异不显著，但 $p < 0.1$，产生了较小的效应量（Cohen, 1969）。以上结果见图 1 所示。

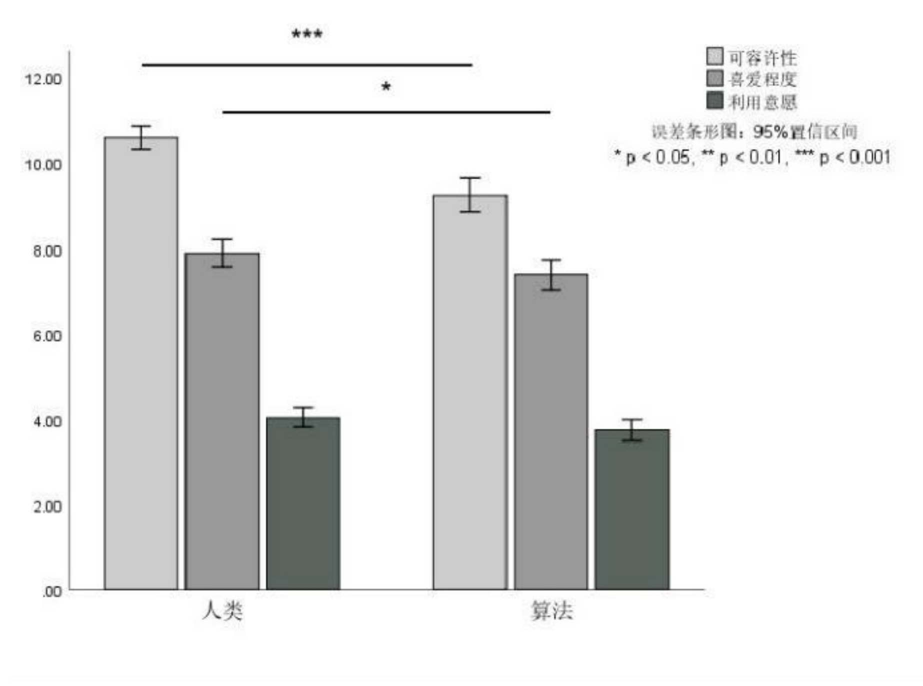



图 1 人类组和算法组的结果比较

注：*p < 0.05, *p < 0.01, ***p < 0.001

以决策主体（人类 vs. 算法）为自变量，反映人们对算法态度的可容许性、喜爱程度、利用意愿三个指标得分为因变量进行多元方差分析（MANOVA）。结果表明，决策主体的主效应显著，Wilks' λ $= 0.896$，$F(3,299) = 11.615$，$p < 0.001$，$\eta_p^2 = 0.104$。

## 2.3 讨论

实验 1 初步验证了在职场中当算法被应用于招聘所做出的决策相较于人类所做的同样决策更不被允许、更不被喜欢，并且人们更不愿意利用算法为自己进行相同任务。该实验证

据基本支持了算法厌恶在职场中的存在。但实验 1 仅仅涉及到人力资源管理工作的一部分，即招聘新员工，且未对算法厌恶背后的机制做进一步的探讨。因此，实验 2 将实验情境设置为应用算法进行年终奖发放工作，以提高实验稳健性，并在其基础上进一步探索感知到的透明性在其中的中介作用。

# 3 实验 2：感知透明性的中介作用

实验 2 在实验 1 的基础上丰富了职场中人力资源管理工作的决策情境，考察了年终奖分配决策，同时进一步探讨感知透明性在其中可能发挥的中介作用。

## 3.1 方法

### 3.1.1 被试

根据 Monte Carlo 模拟法，我们参考实验 1 中效应量 Cohen's $d = 0.63$，取 90%的统计检验力和较窄的稳定性走廊（corridor of stability）宽度 $w = 0.1$，得出实验所需最小样本量为 150（Schönbrodt & Perugini, 2013）。于是，我们通过 Credamo 平台招募被试，实时剔除没有通过注意检查的被试数据并滚动采集，最终得到 179 份有效数据。平均年龄 $M = 30.85$ 岁，$SD = 7.09$ 岁，其中女性 115 名（占比为 64.2%），男性 64 名（占比为 35.8%）。被试被随机分派到人类组（90 人）和算法组（89 人）。所有被试在实验开始之前均仔细阅读了实验说明并知情同意，通过注意检查的有效数据被试在实验结束后获得相应实验报酬。

### 3.1.2 实验设计与程序

同实验 1，实验 2 也采用单因素两水平被试间实验设计，被试首先阅读人力资源决策者（人类或算法）分配年终奖的材料。情境材料改编自 Newman 等人（2020）的研究，两组被试所阅读的材料仅更换了决策主体的身份及相关措辞，而不作其他更改（两段材料的区别以加粗标出）。

人类组阅读："新月公司刚刚完成了年终奖金的发放过程。为了确定每位员工所获得的年终奖数额，新月公司依靠其**资深的人力资源经理张云带领的团队**进行决策，该**团队**考虑了种种因素。在**张云带领的人力资源团队**进行了一系列**审议**过后，确定了员工年终奖的具体分配模式"；

而算法组阅读："新月公司刚刚完成了年终奖金的发放过程。为了确定每位员工所获得的年终奖数额，新月公司依靠**一种可靠的人力资源智能算法"RTC"**进行决策，该**算法**考虑了种种因素。在**智能算法"RTC"**进行了一系列**计算**过后，确定了员工年终奖的具体分配模式"。

在阅读完材料后，被试首先报告自己对上述情境中决策的可容许性、喜爱程度、利用意愿的评分，测量方式同实验 1。可容许性测量在实验 2 中的内部一致性信度为 Cronbach's $\alpha = 0.80$，喜爱程度测量的两题目间相关系数 $r = 0.660$，$p < 0.001$。

随后，我们通过一个题目（参考 Cadario et al., 2021）测量了被试对所阅读材料中决策主体进行决策活动的感知透明性，即“你在多大程度上理解张云（算法“RTC”）是如何做出上述决策的？”，采用李克特 7 点量表计分（从“1 = 完全不能理解”到“7 = 完全能理解”），得分越高代表被试认为所阅读的材料中决策主体的决策更能被理解，亦即感知更透明。

最后，被试回答和实验 1 相同的注意力检查问题（如“此题请选择 1”），这些问题均混杂在测量题目中出现，以便于筛选被试。并且报告了性别和年龄两项人口统计学信息。

### 3.2 结果

#### 3.2.1 决策主体对可容许性、喜欢程度和利用意愿的影响

独立样本 t 检验结果表明，人类组和算法组在三个测量指标上均存在显著差异。以可容许性为因变量时，人类组被试的评分（$M = 12.64$，$SD = 1.89$）显著高于算法组被试的评分（$M = 11.76$，$SD = 2.38$），$t(177) = 2.74$，$p = 0.007$，Cohen's $d = 0.41$；以喜欢程度为因变量时，人类组被试的评分（$M = 11.13$，$SD = 1.98$）显著高于算法组被试的评分（$M = 10.45$，$SD = 2.39$），$t(177) = 2.09$，$p = 0.038$，Cohen's $d = 0.31$；以利用意愿为因变量时，人类组被试的评分（$M = 5.89$，$SD = 0.88$）显著高于算法组被试的评分（$M = 5.27$，$SD = 1.36$），$t(177) = 3.62$，$p < 0.001$，Cohen's $d = 0.54$。

以决策主体（人类 vs. 算法）为自变量，性别和年龄为协变量，反映人们对算法态度的可容许性、喜爱程度、利用意愿三个指标得分为因变量进行多元方差分析（MANOVA）。结果表明，决策主体的主效应显著，Wilks' $\lambda = 0.925$，$F(3,175) = 4.730$，$p = 0.003$，$\eta_p^2 = 0.075$。

#### 3.2.2 感知透明性的中介效应

独立样本 $t$ 检验结果表明，决策主体对感知透明性的影响显著，人类组感知透明性评分（$M = 5.44$，$SD = 1.15$）显著高于算法组（$M = 5.02$，$SD = 1.31$），$t(177) = 2.29$，$p = 0.023$，Cohen's $d = 0.34$。为了进一步探索决策主体对决策态度影响的心理机制，我们使用 Hayes（2013）提供的 SPSS 插件 PROCESS(Model 4)进行探索。我们以决策主体为自变量（人类组 = 1，算法组 = 2），感知透明性为中介变量，分别选取可容许性、喜欢程度和利用意愿为因变量，设定 Bootstrap 样本量为 5000，采用偏差校正的方法，选取 95%置信区间进行中介效应检验。

数据结果显示（如下表所示），当以可容许性为因变量时，感知透明性的中介效应值为–0.50，95%的 Bootstrap 置信区间为[–0.96, –0.08]，不包含 0，表明中介作用显著。并在控制中介变量后，决策主体对可容许性的直接效应为–0.38，95%的 Bootstrap 置信区间为[–0.85, 0.09]，包含 0，表明其直接效应不再显著，证明感知透明性在决策主体对可容许性的影响中起到中介作用。当以喜爱程度为因变量时，感知透明性的中介效应值为–0.56，95%的 Bootstrap 置信区间为[–1.09, –0.09]，不包含 0，表明中介作用显著。并在控制中介变量后，决策主体对可容许性的直接效应为–0.12，95%的 Bootstrap 置信区间为[–0.55, 0.32]，包含 0，表明其直接效应不再显著，证明感知透明性在决策主体对喜爱程度的影响中起到中介作用。当以利用意愿为因变量时，感知透明性的中介效应值为–0.25，95%的 Bootstrap 置信区间为[–0.51, –0.04]，不包含 0，表明中介作用显著。并在控制中介变量后，决策主体对可容许性的直接效应为–0.37，95%的 Bootstrap 置信区间为[–0.63, –0.10]，仍旧不包含 0，表明其直接效应同样显著，感知透明性在决策主体对利用意愿的影响中起到部分中介作用。

表 1 实验 2 中介效应显著性检验的 bootstrap 分析及效应值

| 因变量 | 中介效应值 | 95%间接效应 LLCI | 95%间接效应 ULCI | 直接效应值 | 95%直接效应 LLCI | 95%直接效应 ULCI |
|---|---|---|---|---|---|---|
| 可容许性 | –0.50 | –0.96 | –0.08 | –0.38 | –0.85 | 0.09 |
| 喜欢程度 | –0.56 | –1.09 | –0.09 | –0.12 | –0.55 | 0.32 |
| 利用意愿 | –0.25 | –0.51 | –0.04 | –0.37 | –0.63 | –0.10 |

为了进一步验证中介效应的稳健性，我们又使用传统逐步回归方法进行了中介效应分析（温忠麟 等，2004），结果见下图 2。

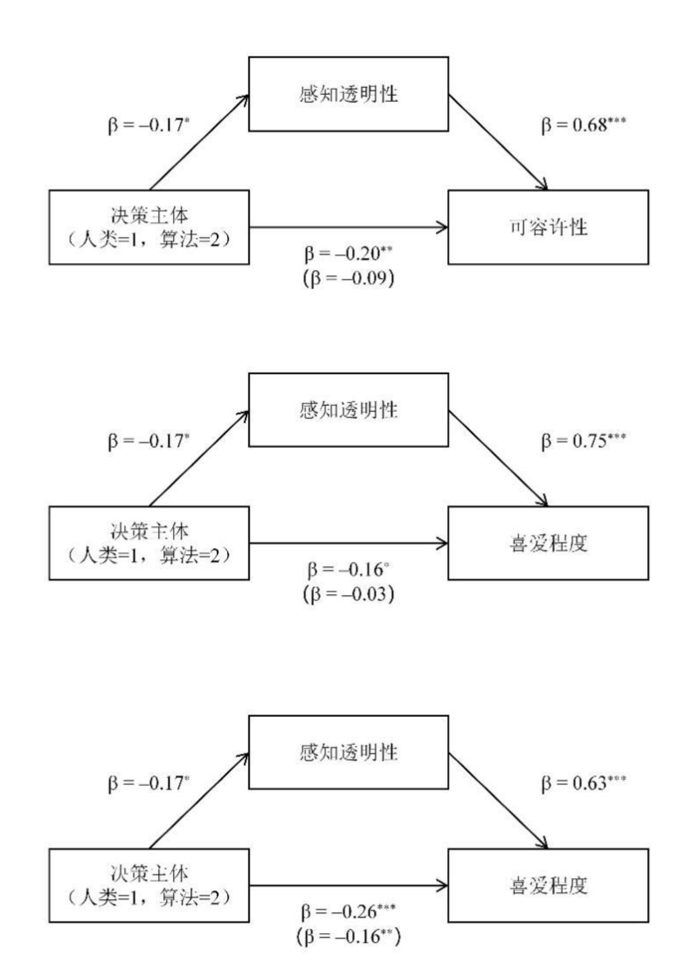



图 2 感知透明性的中介作用

注：*p < 0.05, *p < 0.01, ***p < 0.001

### 3.3 讨论

与实验 1 结果一致，实验 2 再次验证了职场中存在的算法厌恶现象，即人们更偏爱人类做出的决策，更不允许、不喜欢、也不愿再利用算法去做决策，即使两者决策内容完全一致。同时，我们进一步发现了感知透明性在其中发挥的中介作用，即人们厌恶算法 HR 做决策的原因是他们的决策过程被人感知为更不透明，更无法被理解的。实验 1 和实验 2 稳定证明了人们对职场算法的厌恶倾向，同时对感知透明性的中介作用做出了初步探索。为进一步提高对算法厌恶机制的深层次理解，我们将在实验 3 中操纵被试对算法决策的透明性感知，并假设感知透明性越高，被试便越允许、喜欢、愿意利用算法做相关的人力资源决策。

## 4 实验 3：操纵感知透明性

研究 3 我们比较人们对不同透明性算法决策的倾向，并聚焦于简历筛选这一重要的人力资源工作，以提高研究结果稳健性，并进一步深入讨论感知透明性的中介作用。

### 4.1 方法

#### 4.1.1 被试

采用 G*Power 3.1 软件（Faul et al., 2009）计算本实验所需样本量，对于本实验适用的独立样本 $t$ 检验，取中等效应量 $d = 0.5$，显著性水平 $\alpha = 0.05$，计算结果表明，至少需要 172 名被试才能达到 90%统计检验力。通过 Credamo 平台招募被试，实时剔除没有通过注意检查的被试数据并滚动采集，最终得到 180 份有效数据。平均年龄为 31.17±7.41 岁，其中女性 115 名，占比为 63.9%，男性 65 名，占比为 36.1%。被试被随机分派到低透明性组（88 人）和高透明性组（92 人）。所有被试在实验开始之前均仔细阅读了实验说明并知情同意，有效数据被试在实验结束后获得实验报酬。

#### 4.1.2 实验设计与程序

实验为单因素两水平被试间设计。被试被随机分配到高透明性算法和低透明性算法两个组别中的一个。首先，两组被试分别阅读改编自 Newman 等人（2020）研究的算法筛选简历的情境材料。

低透明性组阅读的材料为："在去年的秋季招聘中，冰泉公司使用了一种智能招聘算法"IRA－N"对应聘该公司岗位的毕业生们进行简历评估。智能招聘算法"IRA－N"基于反向传播神经网络进行电子简历的智能筛选。冰泉公司将智能招聘算法所制定的统一格式的电子简历模板通过公司官网发布，并要求所有应聘者均使用此模板进行填写简历。整个简历评估过程，由智能算法"IRA－N"独立完成，全权决定，并对其结果负责。简历筛选评估的过程，是面试工作的基础和重要部分"；

高透明性组阅读的材料为："在去年的秋季招聘中，冰泉公司使用了一种智能招聘算法"IRA－N"对应聘该公司岗位的毕业生们进行简历评估。智能招聘算法"IRA－N"基于反向传播神经网络进行电子简历的智能筛选。**智能算法"IRA－N"将应聘人员的综合素质分成如下五个模块：基本指标（包括学历层次、外语水平、获奖情况等），人格特质（责任心、自信心、包容心等），品行动机（成就动机、学习动机、大局观等），知识技能（学科专业能力、科研成果等），能力素质（岗位工作经验、团队协作能力、创新能力、利用工具能力等）**。冰泉公司将智能招聘算法所制定的统一格式的电子简历模板通过公司官网发布，并要求所有应聘者均使用此模板进行填写简历。**智能招聘算法"IRA－N"已经过大数据 BP 神经网络的深度学习，通过内测，近乎成熟，能够独立且完美的完成电子简历的筛选。**整个简历评估过程，由智能算法"IRA－N"独立完成，全权决定，并对其结果负责。简历筛选评估的过程，是面试工作的基础和重要部分"。

两组被试所阅读的材料中，算法“IRA–N”所面临的任务相同，能力相同，决策结果均未知，但高透明组度材料详细介绍了算法“IRA–N”是根据何种指标做出决策的（以字体加粗标记突出）。由此实现对高低透明性的操纵，即高透明性是指给予被试对算法决策过程和计算能力更多的额外解释。

随后，被试填写实验 2 中测量感知透明性的题项，用于检验我们的操纵是否成功。并填写实验 2 中用于测量可容许性、喜欢程度和利用意愿的题项，在本此次实验中，可容许性三个题目的内部一致性信度 Cronbach’s α = 0.896，喜欢程度两题目相关性显著（$r = 0.854$，$p < 0.001$）。

最后，所有的被试回答与先前实验相同的注意力检查问题，并报告了性别和年龄两项人口统计学信息。

## 4.2 结果

独立样本 $t$ 检验结果表明，高透明性组被试报告的感知透明性得分（$M = 5.68$，$SD = 1.19$）显著高于低透明性组（$M = 4.84$，$SD = 1.68$），$t(178) = 3.90$，$p < 0.001$，Cohen’s $d = 0.58$。说明我们对感知透明性的操纵是有效的。

独立样本 $t$ 检验结果表明，高透明性组被试报告的可容许性评分（$M = 12.25$，$SD = 2.53$）显著高于低透明性组（$M = 10.66$，$SD = 3.42$），$t(178) = 3.56$，$p < 0.001$，Cohen’s $d = 0.53$；高透明性组被试报告的喜欢程度评分（$M = 10.83$，$SD = 2.64$）显著高于低透明性组（$M = 9.05$，$SD = 3.10$），$t(178) = 4.16$，$p < 0.001$，Cohen’s $d = 0.62$；高透明性组被试报告的利用意愿评分（$M = 5.57$，$SD = 1.33$）显著高于低透明性组（$M = 4.68$，$SD = 1.66$），$t(178) = 3.94$，$p < 0.001$，Cohen’s $d = 0.59$。上述结果如下图 3 所示。

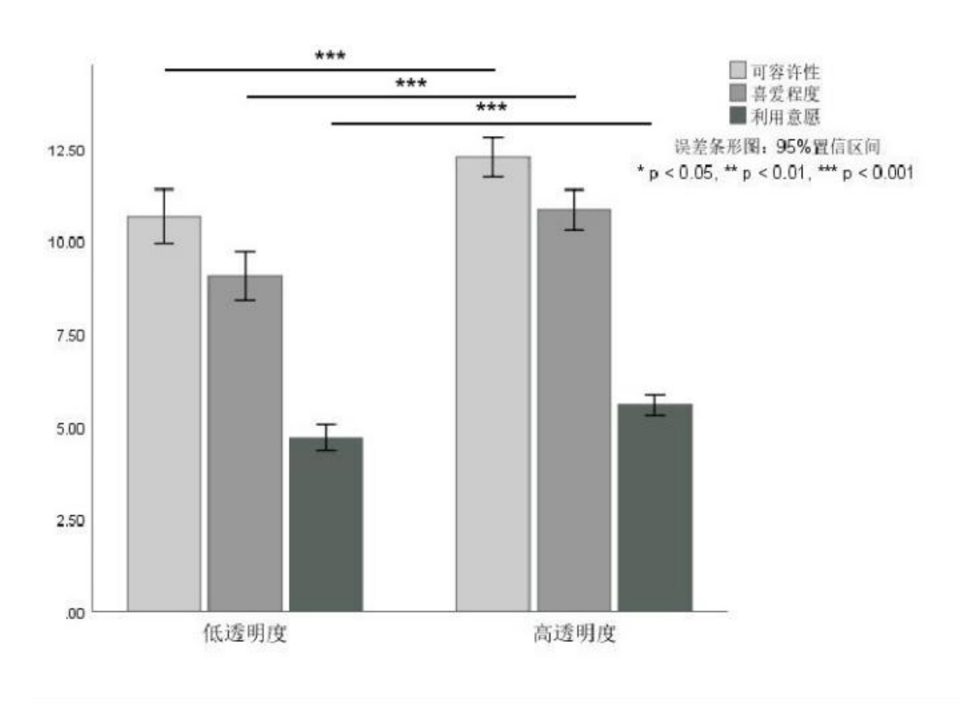



图 3 不同透明度条件的结果比较

注：*p < 0.05, **p < 0.01, ***p < 0.001

以透明性为自变量，性别和年龄为协变量，反映人们对算法态度的可容许性、喜爱程度、利用意愿三个指标得分为因变量进行多元方差分析（MANOVA）。结果表明，透明性的主效应显著，Wilks' λ = 0.907，$F(3,176) = 6.035$，$p = 0.001$，$\eta_p^2 = 0.093$。

4.3 讨论

实验 3 通过直接操纵人们对算法决策感知透明性，发现提高人们对算法决策的感知透明性能够显著地提高人们对算法决策的可容许性、喜欢程度和利用意愿，进一步验证了感知透明性这一变量在职场算法厌恶中发挥的中介作用。既然人们是因为无法理解算法，认为算法是一个黑箱而厌恶算法 HR 做出的决策，那么当算法被拟人化时，使本不是人的算法更加像人是否会对算法厌恶的产生起到调节作用，并降低这种倾向性呢？为了回答这一问题，并继续探索职场算法厌恶的边界条件，我们将通过实验 4 来验证算法拟人化在其中可能起到的调节作用。

# 5 实验 4：算法拟人化的调节作用

为了探讨上述实验所证明出的职场中的算法歧视存在怎样的边界条件，研究 4 考察了算法拟人化对算法厌恶的调节作用。此外，在实验 4 中我们着眼于人力资源管理中绩效考核的工作，从而进一步丰富了研究所涵摄的人力资源工作内容。

5.1 方法

5.1.1 被试

基于实验 4 的单因素三水平被试间设计，使用 G*Power3.1 软件（Faul et al., 2009）计算本实验所需最少样本量，在显著性水平 α = 0.05 且中等效应量（*f* = 0.25）时，预测达到 90% 的统计检验力水平总共需要至少 207 名被试。为保证实验数据的稳健性，本实验通过 Credamo 和 PowerCX 招募被试，随机分配到非拟人化算法组、拟人化算法组和人类组，平台实时剔除未通过注意力检查的被试，最终剩余有效样本 549 名（其中女性为 327 名，占 59.6%）。被试年龄 *M* = 31.4 岁，*SD* = 7.88 岁。其中非拟人化算法组共 181 名被试，拟人化算法组共 185 名被试，人类组 183 名被试。所有被试在实验开始之前均仔细阅读了实验说明并知情同意，有效数据被试在实验结束后获得了一定数额的报酬。

#### 5.1.2 实验设计与程序

实验 4 为单因素三水平被试间实验设计，三组分别为人类组、拟人化算法组以及非拟人化算法组，所有被试被随机分配到其中一组。首先，被试被要求阅读了不同决策主体进行绩效考核的情境材料。

人类组阅读的情境材料为："您好，我是黄飞，我是初禾集团的人力资源经理。本次初禾集团的员工绩效考核由我负责进行。我会评估每位员工的绩效数据。此外，我还会对员工录制的 5 分钟自我陈述视频进行评估。在评估绩效数据和视频之后，我会对本年度员工的绩效考核做出最终决定，独立完成全权负责。这项决定结果可能会影响到员工的工资、奖金、晋升资格，甚至在某些情况下会影响是否解聘该员工"。

拟人化算法组阅读的情境材料为："嗨，你好!我叫阿奇，我是您的人力资源智能助手。本次初禾集团的员工绩效考核由我负责进行。我会评估每位员工的绩效数据。此外，我还会对员工录制的 5 分钟自我陈述视频进行评估。在评估绩效数据和视频之后，我会对本年度员工的绩效考核做出最终决定，独立完成，全权负责。这项决定结果可能会影响到员工的工资、奖金、晋升资格，甚至在某些情况下会影响是否解聘该员工"。

非拟人化算法组阅读的情境材料为：""HRA"是一种新型的人力资源智能算法。本次初禾集团的员工绩效考核由算法"HRA"负责进行。算法"HRA"会评估每位员工的绩效数据。此外，算法"HRA"还会对员工录制的 5 分钟自我陈述视频进行评估。在评估绩效数据和视频之后，人力资源智能算法"HRA"会对本年度员工的绩效考核做出最终决定，独立完成，全权负责。这项决定结果可能会影响到员工的工资.奖金、晋升资格，甚至在某些情况下会影响是否解聘该员工"。

作为对算法拟人化的操纵，拟人化算法组与非拟人化算法组的差别在于改变其表达方式的叙事视角，即拟人化算法组采用第一人称的主观陈述方式、非拟人化算法组采用第三人称

的客观陈述方式。该操纵参考了已有研究对拟人化程度的操纵范式（Hur et al., 2015; May & Monga, 2014），即为非人对象起一个人名，并以第一人称加以描述，能够有效提升拟人化程度。除此以外，两组对于算法的描述完全相同。并且，为了检验拟人化操纵的有效性，拟人化算法与非拟人化算法组被试均需在回答其他问题之前，完成对情境中算法的拟人化程度的评分（你认为人力资源智能助手“阿奇”/智能算法“HRA”在多大程度上让你想起了一些人类的特质？，该测量采用李克特 7 点量表计分，从“1＝ 一点也没有”到“7＝ 非常多”)，该操纵检查改编自 Hur 等人（2015）的研究。

在阅读完上述材料并进行操纵检查后，三组被试分别报告了对其所阅读到的决策主体所做决策的可容许性、喜爱程度和利用意愿，测量条目同上述实验。

最后，被试回答了同之前实验的注意力检查问题，用于筛选被试，并报告了性别、年龄、对算法的熟悉程度（“你对人工智能算法有多熟悉”，从“1＝ 完全不熟悉”到“7＝ 非常熟悉”）和了解程度（“你对人工智能算法有多了解”，从“1＝ 完全不了解”到“7＝ 非常了解”，两题均参考自 Leo 和 Huh（2020）的研究）四项人口统计学信息。

5.2 结果

独立样本 t 检验结果表明，对拟人化算法的拟人化评分（$M = 5.46$，$SD = 1.23$）显著高于对非拟人化算法的拟人化评分（$M = 4.87$，$SD = 1.45$），$t(364) = –4.32$，$p < 0.001$，Cohen's $d = 0.44$。说明本实验对于算法拟人化的操作是有效的。

以可容许性作为因变量进行单因素方差分析发现，决策主体的主效应显著，$F(2,546) = 3.15$，$p = 0.044$，$\eta_p^2 = 0.11$；以喜爱程度作为因变量进行单因素方差分析发现，决策主体的主效应不显著，$F(2,546) = 2.39$，$p = 0.093$，$\eta_p^2 = 0.09$；以利用意愿作为因变量进行单因素方差分析发现，决策主体的主效应不显著，$F(2,546) = 0.40$，$p = 0.668$。

计划对比（planned contrast）分析表明，当以可容许性作为因变量时，人类组得分（$M = 10.54$，$SD = 2.81$，$p = 0.021$，Cohen's $d = 0.25$）显著高于非拟人化组算法得分（$M = 9.81$，$SD = 3.11$），而拟人化算法组得分（$M = 10.42$，$SD = 2.94$）在数值上高于非拟人化算法组得分（$M = 9.81$，$SD = 3.11$，$p = 0.055$，Cohen's $d = 0.20$），差异不显著但表现出接近统计学意义的显著性水平，并产生较小的效应量（Cohen, 1969），另外，人类组与拟人化算法组之间并无差异，$p = 0.705$；当以喜爱程度作为因变量时，人类组平均得分（$M = 9.13$，$SD = 3.16$）虽高于非拟人化算法组平均得分（$M = 8.68$，$SD = 3.43$），但未达到统计显著性标准，$p = 0.192$，没有表现出具有统计学意义的算法厌恶。而拟人化算法组得分（$M = 9.40$，$SD = 2.94$，$p = 0.032$，Cohen's $d = 0.23$）显著高于非拟人化算法组得分（$M = 8.68$，$SD = 3.43$），

表明算法拟人化对于提高人们对算法决策的喜爱程度是有作用的。上述结果如下图 4 所示。

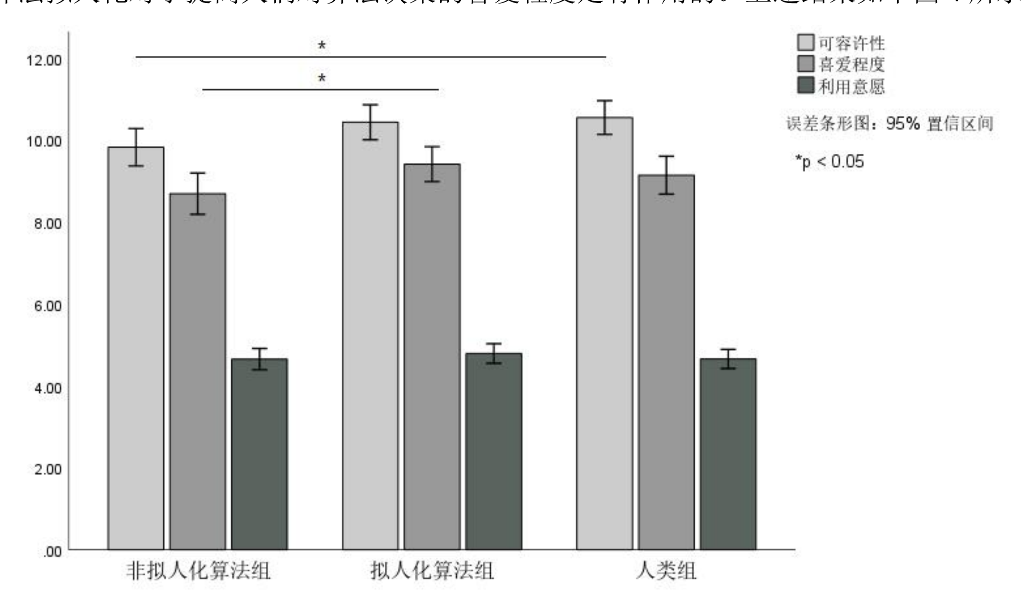



图 4 不同决策主体组的结果比较

注：*p < 0.05

以不同拟人化程度的决策主体（人类 vs. 拟人化算法 vs. 非拟人化算法）为自变量，以对算法熟悉程度、对算法了解程度为协变量，以反映人们对算法态度的可容许性、喜爱程度、利用意愿三个指标的得分为因变量进行多元方差分析（MANOVA）。结果表明，决策主体的主效应显著，Wilks' λ = 0.965，$F(6,1084) = 3.251$，$p = 0.004$，$\eta_p^2 = 0.018$。

当将被试报告的对算法的熟悉、了解程度作为协变量，以组别为自变量，当以可容许性为因变量时，进行方差分析，其结果显示，对算法熟悉程度：$F(1,544) = 4.16$，$p = 0.042$，$\eta_p^2 = 0.008$；对算法了解程度：$F(1,544) = 5.64$，$p = 0.018$，$\eta_p^2 = 0.010$，对算法的熟悉、了解程度的效应显著；组别的效应差异不显著但十分接近显著性标准，并产生较小的效应量，$F(2,544) = 2.98$，$p = 0.052$，$\eta_p^2 = 0.011$。保持自变量、协变量不变，当以喜爱程度为因变量时，结果显示，对算法熟悉程度：$F(1,544) = 9.45$，$p = 0.002$，$\eta_p^2 = 0.017$；对算法了解程度：$F(1,544) = 9.67$，$p = 0.002$，$\eta_p^2 = 0.017$；组别的效应同样接近显著性标准，并产生较小的效应量，$F(2,544) = 2.62$，$p = 0.074$，$\eta_p^2 = 0.01$。由此可知，关于算法的先验性的熟悉、了解程度对于人们对待算法管理的态度有极大的影响，这与前人研究保持一致（Ireland, 2020; Komatsu, 2016）。这说明当人们对算法越加熟悉和了解，就越会接纳算法在职场中的各种决策应用，越赋予算法决策以合法地位。换言之，这说明提升民众的算法意识和算法素养同样有助于改善人们的厌恶倾向。

综上所述，算法拟人化在决策主体对可容许性和喜爱程度的影响中具有一定的调节作用。

### 5.3 讨论

实验 4 证明了算法拟人化在决策主体对决策态度影响中的调节作用。具体而言，将算法拟人化能够显著地提高人们对算法决策的喜爱程度，表达出更少的算法厌恶。但拟人化并不能有效提高人们对算法决策者的利用意愿。

## 6 总讨论

本研究探索了在职场应用场景下，人们对人类决策者和算法决策者的态度是否存在差异，并在此基础上探讨了造成如此差异的内在机制和边界条件。通过 4 项递进实验，我们发现相对于人类决策者，当算法走上人力资源管理岗位时会引发人们更加苛刻的评价，这是由于算法决策（相比于人类）通常被人们感知为是更不透明的，并且这一差异受到算法拟人化的调节。具体而言，通过为不同被试呈现由人类或算法完成的同样的人力决策并测量其态度，我们发现人们更不允许、更不喜欢、更不愿意利用算法所做的决策，在认知、情感、行为三个不同维度上均会表现出对算法决策排斥，且这一差异具有一定程度的稳健性（实验 1~4）。通过测量被试感知到的决策透明性（实验 2）以及操纵算法决策的可解释性（实验 3），我们进一步发现感知透明性是造成人们对不同决策主体（人类 vs.算法）产生不同决策态度的心理机制，即相比于人类决策，算法进行同样的决策时，人们认为其决策过程是更不透明、更不可理解的，因此造成了对算法的回避。通过对算法拟人化程度的操纵（实验 4），我们发现了决策主体对决策态度的影响受到拟人化的调节作用。当算法拟人化程度高时，人们对算法决策的态度便会更加宽容。这说明拟人化算法是减少职场中算法厌恶的有效途径之一。并且在分析控制变量的过程中，我们发现了被试对算法的熟悉、了解程度的个体差异能够正向预测其对算法管理的积极态度，这似乎说明提升民众的算法素养也能有效推进智能管理。在研究中我们考察了算法决策在职场中的不同应用场景，包括招聘录用（实验 1）、年终奖分配（实验 2）、简历筛选（实验 3）、绩效考核（实验 4）；并且研究的样本涵盖了不同被试，包括来自 Credamo（实验 2~4）、PowerCX（实验 4）的全国范围内被试以及来自某高校的大学生被试（实验 1）。正因此，多样化的实验情境材料和被试选取保障了研究结果的稳健性。

### 6.1 被拒斥的算法管理

随着工业 4.0 时代的逐步推进，由信息技术赋能的算法越发深入地介入到人们社会生活的方方面面之中，人与自动化机器（泛指计算机算法、人工智能等等）的关系发生了革命性的范式转变。从原本类似主奴关系的“用户-工具（user-tool）”范式，进展到更为平等的“合

作伙伴（partner）”范式，到如今方兴未艾的“下属-领导（subordinate-leader）”关系模式彻底颠覆了原先人们对人与机器关系的认识和理解（Wesche & Sonderegger, 2019）。经典的计算机作为社会行动者（computers as social actors, CASA）范式认为，人们会将计算机和其他先进信息技术视作独立的社会实体，与之的互动也会遵从人类社会的社交规则，而不简单认为其是人类编程的死板呈现（Nass et al., 1994）。由此引申出的“计算机作为领导者（computers as leaders）”范式强调，自动化算法将成为管理层级结构中的中层领导，负责实现高层管理者与底层员工之间的上传下达（Wesche & Sonderegger, 2019）。在此背景下，对于人们如何看待算法管理的研究是大有裨益的。本研究发现了人们对职场中相关决策的算法使用持较为稳定的排斥和厌恶态度，这与先前研究结果基本一致，无论是普通民众还是管理者都对算法自主决策感到不安和排斥（Acikgoz et al., 2020; Diab et al., 2011; Fischer & Peterson, 2018; Haesevoets et al., 2021; Nørskov et al., 2020）。本研究发现，当人类与算法做出相同的人力决策时，人们依然会更抵触算法管理。这一结论为职场中的算法厌恶提供了新的证据。同时，也提醒研究者和企业管理者们需要重审作为领导的算法与人类之间的关系，并思考如何使算法管理在普通员工的心理层面平滑过渡，如何让人力算法真正被接纳为企业中的一个独立主体。

值得一提的是，本研究通过态度的不同维度以及职场中的不同视角构建出“三维度-两视角”的因变量模型以测定人们对职场中算法使用的厌恶倾向，扩充了先前研究在表征算法厌恶上的片面性。过往研究大多都只测量被试对算法决策或机器人服务的一个反应指标，例如信任（Hoff & Bashir, 2015）、购买意愿（Wien & Peluso, 2021）等。但这显然是单薄的，人们会出现态度中知行的不一致，比如认可算法优越性的同时排斥使用算法推荐系统（Yeomans et al., 2019）。结合对人类心理活动的认识，我们认为算法厌恶需要从认知、情感、行为（意图）三个维度加以理解和考察（参考态度 ABC 理论，Breckler, 1984）。因此，本研究选取了可容许性、喜爱程度、利用意愿三个变量作为上述三个维度的表征。可容许性从认知维度反映了人们对决策主体进行决策活动的合法性认识，即一般地认为某一决策主体具有行使决策的能力、权利、资格；喜爱是人接应外物而产生的一种积极情感，也是人能在短时间内对外物做出的直觉判断（Bartneck et al., 2009），其反映了人们在情感上对所面对之事物的倾向性；利用意愿则考察了被试基于假想的管理者视角的行为意图。另外，为力求立体还原，本研究所选取的因变量还兼顾职场中“员工-领导（employee-employer）”的双重视角（Cummins, 1998; Gigerenzer & Hug, 1992）。其中可容许性和喜爱程度是从员工的角度来看待算法，而利用意愿则是要求被试想象自己作为企业负责人对算法的接受性反应。这

一视角上的差异或许说明了为什么被试在利用意愿上的算法厌恶在实验1和实验4中并没有得到复制。或许原因在于被试样本的年轻化（其中很大一部分为学生被试），导致其并没有作为企业负责人的生活经验和体会，甚至可能没有与企业负责人交际的经历，很难想象这一身份的选择倾向性。

诚然，本研究结论证明人们存在一种对职场中算法使用的普遍厌恶倾向，但这并不意味着企业因此就要放弃数字化进程。原本为提升员工福祉和企业绩效而发明设计的算法技术（Benlian et al., 2022），虽不幸因其不透明的属性而从解放人反过来成为奴役人的超级工具，最终招致怀疑和拒斥（Jussupow et al., 2020）。但这不足以让人类选择一条因噎废食的错误道路。并且，在对算法管理的优化升级过程中，研究者和一线管理人员不仅需要琢磨如何设计出更为人接纳、助人成就的算法决策系统（Simth & Shum, 2018），仍然需要关注在算法管理背景之下人机关系、人人关系会出现何种程度的变化。总而言之，本文实际上是在敲响警钟，面对推动管理革新的算法技术，我们必须正确地审视其“利弊兼具”的双刃剑特性，扬长而避短，发挥出其最大效能。

### 6.2 通向智能管理的可行之路

从实践价值上来说，本研究致力于探索接纳算法管理的关键因素，从而推动人力决策的自动化、智能化。结合四个子实验的发现，本研究提出三条通向智能管理的可行道路。

第一，提高算法决策的透明性，开发可解释性人工智能。实验2发现，人们对算法和人类两种不同决策主体的态度差异根源在于人们认为算法的决策过程相较于人类更加不透明、更加不可理解。实验3发现，当打开“黑箱”，提升算法决策的可解释性，能够有效改善人们对其原本抱持的消极态度。总而言之，透明性作为中介机制解释了人们对算法管理的厌恶态度，同时暗示出一个提升智能管理效能的可行方案。一般而言，算法的透明性包含两方面意指，其一是指作为决策主体的算法呈现其自身决策的过程和依据，使原本不可知的过程可知，另一方面则是向观测者展现算法运算的底层规则和逻辑（Confalonieri et al., 2021; Leichtmann et al., 2023）。因此，所谓提升算法透明性，在技术层面须通过开发可解释性人工智能（Explainable Artificial Intelligence, XAI），使得使用者能够理解、清楚AI算法做决策之原委，使原本密不透光之“黑箱（black box）”转变为澄明剔透之“玻璃箱（glass box）”（Rai, 2019）。

第二，设计拟人化的管理类算法。实验4发现，对算法进行姓名、表达风格的浅层拟人化处理能够有效改善人们对其的厌恶态度，即相比于拥有机械名字、第三人称表述的算法管理者，人们更容许、更喜欢由拥有类人名字、第一人称表述的算法管理者做出绩效考核决策。

这与先前关于拟人化带来优势效应和积极体验的研究基本保持一致（Han, 2021; Natarajan & Gombolay, 2020; Yuan & Dennis, 2019）。基于研究结论，本研究倡导算法管理系统的设计方可采用拟人化的形式，将“冷酷无情”的算法升温，以提高人们对其进入决策领域的接纳程度。

第三，提高民众的算法素养。在实验 4 中，本研究发现对算法的熟悉和了解程度两个控制变量对因变量有正向的影响，即对算法越熟悉越了解的被试，其对算法管理的态度也更为积极。这一发现可以用简单暴露效应（mere exposure effect）进行解释，即只要将某些外部信息反复呈现给人，人们对其的喜爱程度就有可能会提高（Zajonc, 1968）。这发现说明了如果想要提高算法管理在群众之中的接受程度，或许可以通过提升人们对算法的熟悉和了解程度。更广义而言，则需要提高民众的算法素养（algorithmic literacy），即用户围绕算法产生的意识、知识、想象、策略和技能（Swart, 2021）。生活于自动化、信息化、智能化的社会中，人们需要提高自身的知识见闻以应对须臾不可离的算法。同时告知管理者们应当致力于培养员工的算法意识、算法身份（即 IT identity，Craig et al., 2019），从而能够更好地推动现代管理的智能化。

## 6.3 研究局限与未来展望

当然，本研究仍存在一定的局限性。第一，在因变量选取上虽然涵盖了人们对算法反应的认知、情感、行为三维度和管理者-下属两个视角，但也只是关注了每个维度其中的一个指标，仍有许多其他的重要因素有待后续实验测量：例如信任（Logg et al., 2019）、公平感（Schoeffer et al., 2022）、恐怖感（Mende et al., 2019）、道德指责（Malle et al., 2016）、惩罚行为（Lokhorst & van den Hoven, 2011）等等。另外，利用意愿仅仅能代表人们自我报告出的对再次使用该种决策者的行为意图，缺少相对客观和准确的测量，不具有较好的生态效度。并且，人们很可能受困于口是心非、知行不一，在真实的管理场景中做出与实验情境中不相一致的反应。因此，未来研究可以将本研究揭示的效应放入真实的组织环境中再做验证，采用现场观察等方式记录被试对不同类型管理者（涉及人与不同形式的算法）的真实反映。

其次，本研究仅仅证明了赋予算法以人类名称和拟人化的言辞表达（例如，第一人称）这种最表层的拟人化方法的有效性。一方面这是因为言辞表达和名称使用是实际应用中最为简单的拟人化方法（例如苹果公司的智能语音助手 Siri、小米公司的语音交互引擎小爱同学早已投入生产和使用），最能够得到普遍的应用；另一方面也是因为外观、动作等更加深度的拟人化对人们态度的影响更加复杂，甚至可能出现恐怖谷效应（uncanny valley effect,

Laakasuo et al., 2021; Mori, 1970; Mori et al., 2012）、身份威胁（Yoggeeswaran et al., 2016; Zlotowski et al., 2017）等适得其反的干扰。因此，未来研究可以更加深入探讨不同类型、不同程度的拟人化能对人们对于职场算法的态度产生如何的影响。

当然，职场中算法厌恶可能仍存在其他解释机制和边界条件。在本研究中，我们着重探究了感知透明性的中介作用。可实际上人们的认知是复杂多维的，其他因素也可能成为中介。例如，更少的自由意志（许丽颖 等, 2022）、更低的心智感知水平（Bigman & Gray, 2018）、独特性威胁（Ferrari et al., 2016）都可能是职场决策中算法厌恶的潜在原因。未来的研究可以更加细致地去考察这些可能的变量，并对这些影响因素加以综合比较，以便更彻底地认识人对算法与对人类的反应差异的复杂机制。同样，影响人们的算法接纳度的边界条件绝不止拟人化。对于机器接纳度的探索和降低算法厌恶的努力至少可以从三方面来理解，即人类自身特质、机器自有属性和人机互动模式（许丽颖, 喻丰, 2020）。因此，算法使用者的个体差异、算法及其实体给使用者造成的不同心理感知（如温暖/能力，Fiske et al., 2002）、人-算法协作的权重模式都可能会潜在地影响对算法管理接纳度，这仍有待后续研究不断深入探索。

# 7 结论

本研究结论如下：第一，在职场诸应用场景中，相对于人类决策者，人们对算法决策的容许、喜爱、利用程度均更低，表现出稳定地算法厌恶倾向；第二，这一现象的内在心理机制是人们认为算法（相比于人类）的决策更难理解、更不透明；第三，算法越是拟人化，人们对算法决策的厌恶倾向越低。

# 参考文献


Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment, 28*(4), 399–416.

Adam, M., Wessel, M., & Benlian, A. (2021). AI–based chatbots in customer service and their effects on user compliance. *Electronic Markets, 31*(2), 427–445.

Ahmed, S., Alshater, M., Ammari, A., & Hammami, H. (2022). Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance, 61*, 101646.

Andrews, D., Bonta, J., & Wormith, J. (2006). The recent past and near future of risk and/or need assessment. *Crime and Delinquency, 52*(1), 7–27.

Badue, C., Guidolini, R., Carneiro, R., Azevedo, P., Cardoso, V., Forechi, A., Jesus, L., Berriel, R., Paixão, T. M., Mutz, F., de Paula Veronese, L., Oliveira-Santos, T., & De Souza, A. (2021). Self–driving cars: A survey. *Expert Systems with Applications, 165*, 113816.

Barnes, C. M., Lucianetti, L., Bhave, D. P., & Christian, M. S. (2015). "You wouldn't like me when I'm sleepy": Leaders' sleep, daily abusive supervision, and work unit engagement. *Academy of Management Journal, 58*(5), 1419–1437.

Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71–81.

Basu, S., Majumdar, B., Mukherjee, K., Munjal, S., & Palaksha, C. (2023). The role of artificial intelligence in HRM: A systematic review and future research direction. *Human Resource Management Review, 33*(1),100893.

Benlian, A., Wiener, M., Cram, W., Krasnova, H., Maedche, A., Möhlmann, M., Recker, J., & Remus, U. (2022). Algorithmic management: Bright and dark sides, practical implications, and research opportunities. *Business & Information Systems Engineering, 64*(6), 825–839.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34.

Blair, A., & Saffidine, A. (2019). AI surpasses humans at six–player poker. *Science, 365*(6456), 864–865.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.

Bostrom, N. (2014). *Superintelligence*. New York: Oxford University Press.

Breckler, S. J. (1984). Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of Personality and Social Psychology, 47*(6), 1191–1205.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age : Work, progress, and prosperity in a time of brilliant technologies*. New York : W.W. Norton & Company.

Burrell, J. (2016). How the machine'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 205395171562251.

Burton, J., Stein, M., & Jensen, T. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239.

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour, 5*(12), 1636–1642.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *The American Economic Review, 106*(5), 124–127.

Chen, Z., Liu, Y., Meng, J., & Wang, Z. (2023). What's in a face? An experiment on facial information and loan-approval decision. *Management Science, 69*(4), 2263–2283.

Cheng, M., & Hackett, R. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review, 31*(1), 100698.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 11*(1), e1391.

Craig, K., Thatcher, J. B., Grover, V. (2019). The IT identity threat: A conceptual definition and operational measure. *Journal of Management Information Systems, 36*(1), 259–288.

Cummins, D. (1998). Social norms and other minds: The evolutionary roots of higher cognition. In D. D. Cummins & C. Allen (Eds), *The evolution of mind* (pp. 30–50). New York: Oxford University Press.

Dane, E., Rockmann, K. W., & Pratt, M. (2012). When should I trust my gut? Linking domain expertise to intuitive decision–making effectiveness. *Organizational Behavior and Human Decision Processes, 119*(2), 187–194.

De Visser, E., Monfort, S., McKendrick, R., Smith, M., McKnight, P., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*(3), 331–349.

Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non–US samples. *International Journal of Selection and Assessment, 19*(2), 209–216.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.

Dodge, J., Liao, Q., Zhang, Y., Bellamy, R., & Dugan, C. (2019, March). Explaining models: An empirical study of how explanations impact fairness judgment. Paper presented at the meeting of 24th ACM International Conference on Intelligent User Interfaces, Marina del Ray, United State.

Duggan, J., Sherman, U., Carbery, R., & McDonnell, A. (2020). Algorithmic management and app–work in the gig economy: A research agenda for employment relations and HRM. *Human Resource Management Journal, 30*(1), 114–132.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three–factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

Ferrari, F., Paladino, M., & Jetten, J. (2016). Blurring human–machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics, 8*(2), 287–302.

Filiz, I., Judek, J., Lorenz, M., & Spiwoks, M. (2021). Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance, 31*, 100524.

Fischer, S., & Peterson, T. (2018). *Was Deutschland über Algorithmen weiß und denkt: Ergebnisse einer repräsentativen Bevölkerungsumfrage [What Germany knows and think about algorithms: Results of a representative survey]*. Gütersloh, Germany: Bertelsmann Stiftung.

Fiske, S. T., Cuddy, A. J. C., Glick, P., Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

Fraune, M. R. (2020). Our robots, our team: Robot anthropomorphism moderates group effects in human–robot teams. *Frontiers in Psychology, 11*, Article 1275.

Garg, S., Sinha, S., Kar, A., & Mani, M. (2022). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management, 71*(5), 1590–1610.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating and perspective change. *Cognition, 43*(2), 127–171.

Goods, C., Veen, A., & Barratt, T. (2019). "Is your gig any good?" Analysing job quality in the Australian platform–based food–delivery sector. *Journal of Industrial Relations, 61*(4), 502–527.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124.

Haesevoets, T., De Cremer, D., Dierckx, K., & Van Hiel, A. (2021). Human-machine collaboration in managerial decision making. *Computers in Human Behavior, 119*, 106730.

Han, M. (2021). The impact of anthropomorphism on consumers' purchase decision in chatbot commerce. *Journal of Internet Commerce, 20*(1), 46–65.

Hao, K. (2019). AI is sending people to jail—and getting it wrong. MIT *Technology Review,* Retrieved January 21, 2019, from https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

Hao, K. (2020). Doctors are using AI to triage covid–19 patients. The tools may be here to stay. *MIT Technology.*

*Review,* Retrieved April 23, 2020, from https://www.technologyreview.com/2020/04/23/1000410/ai-triage-covid-19-patients-health-care/

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression–based approach*. New York: Guilford Press.

Hays, K. (2022). Facebook contractors learned they lost work with the company through a video call with anonymous representatives who said an ‘algorithm’ chose random people to cut, workers say. *Business Insider*. Retrieved August 19,2022,from https://www.businessinsider.com/facebook-contract-workers-accenture-austin-lost-jobs-2022-8

Herlocker, J., Konstan, J., Terveen, L., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5–53.

Hitsuwari, J., Ueda, Y., Yun, W., & Nomura, M. (2023). Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human–made and AI–generated haiku poetry. *Computers in Human Behavior, 139*, 107502.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(3),407–434.

Holford, W. (2022). An ethical inquiry of the effect of cockpit automation on the responsibilities of airline pilots: Dissonance or meaningful control? *Journal of Business Ethics, 176*(1), 141–157.

HR Daily Advisor Staff. (2017). Artificial Intelligence will become a regular part of HR in next 5 years. *HR Daily Advisor*. Retrieved June 8, 2017, from https://hrdailyadvisor.blr.com/2017/06/08/artificial-intelligence-will-become-regular-part-hr-next-5-years/

Hur, J. D., Koo, M., & Hofmann, M. (2015). When temptations come alive: How anthropomorphism undermines self–control. *Journal of Consumer Research, 42*(2), 340–358.

Institute of Electrical and Electronics Engineers (IEEE). (2017). *Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems*. Author

Ireland, L. (2020). Who errs? Algorithm aversion, the source of judicial error, and public support for self-help behaviors. *Journal of Crime and Justice, 43*(2), 174–192.

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries, 5*(1), 38–56.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020, June). W*hy are we averse towards algorithms? A comprehensive literature review on algorithm aversion*. Paper presented at the meeting of the Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference, Marrakech, Morocco.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review, 93*(5), 1449–1475.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Kharpal,A. (2023). A.I. poses existential risk of people being 'harmed or killed,' ex-Google CEO Eric Schmidt says. *Consumer News and Business Channel*. Retrieved May 24, 2023, from https://www.cnbc.com/2023/05/24/ai-poses-existential-risk-former-google-ceo-eric-schmidt-says.html

Kinowska, H., & Sienkiewicz, Ł. J. (2022). Influence of algorithmic management practices on workplace well-being–evidence from European organisations. *Information Technology & People (West Linn, Or.), 36*(8), 21–42.

Komatsu, T. (2016, March). *Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds*. Paper presented at the meeting of 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633–705.

Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics, 13*(7), 1679–1688.

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment, 27*(3), 217–234.

Langer, M., König, C., & Busch, V. (2021). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology, 36*(5), 751–769.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 2053951718756684.

Lee, T., & Boynton, L. A. (2017). Conceptualizing transparency: Propositions for the integration of situational factors and stakeholders' perspectives. *Public Relations Inquiry, 6*(3), 233–251.

Lehdonvirta, V. (2018). Flexibility in the gig economy: Managing time on three online piecework platforms. *New Technology, Work, and Employment, 33*(1), 13–29.

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior, 139*, 107539.

Lemieux, P. (2017). Rise of the machines? *Regulation, 40*(4), 4–5.

Leo, X., & Huh, Y. (2020). Who gets the blame for service failures? Attribution of responsibility toward robot versus human service providers and service firms. *Computers in Human Behavior, 113*, 106520.

Lerner, J., Li, Y., Valdesolo, P., & Kassam, K. (2015). Emotion and decision making. *Annual Review of Psychology, 66*(1), 799–823.

Li, X., & Sung, Y. (2021). Anthropomorphism brings us closer: The mediating role of psychological distance in User–AI assistant interactions. *Computers in Human Behavior, 118*, 106680.

Liu, N., Kirshner, S., & Lim, E. (2023). Is algorithm aversion WEIRD? A cross-country comparison of individual-differences and algorithm aversion. *Journal of Retailing and Consumer Services, 72*, 103259.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes,151*, 90–103.

Lokhorst, G. J., & van den Hoven, J. (2011). Responsibility for military robots. In P. Lin, K. Abeney, & George A. Bekey (Eds.). *Robot ethics: the ethical and social implications of robotics* (pp. 145–156). Cambridge, MA: The MIT Press.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650.

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The “word‐of‐machine” effect. *Journal of Marketing, 86*(1), 91–108.

Mahmud, H., Islam, A., Ahmed, S., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting & Social Change, 175*, 121390.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March). *Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot*. Paper presented at the meeting of 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand.

May, F., & Monga, A. (2014). When time has a will of its own, the powerless don't have the will to wait: Anthropomorphism of time can decrease patience. *Journal of Consumer Research, 40*(5), 924–942.

McAloon, A. (2021). Xsolla lays off 150 after an algorithm ruled staff ‘unengaged and unproductive’. *Game Developer*. Retrieved August 10, 2021, from https://www.gamedeveloper.com/business/xsolla-lays-off-150-after-an-algorithm-ruled-staff-unengaged-and-unproductive-#close-modal

McFarland, M. (2014). Elon Musk: ‘With artificial intelligence we are summoning the demon.’ *The Washington*

*Post*. Retrieved October 24, 2014, from https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/

McNee, S., Riedl, J., & Konstan, J. (2006, April). *Being accurate is not enough: How accuracy metrics have hurt recommender systems*. Paper presented at the meeting of CHI '06 Extended Abstracts on Human Factors in Computing Systems, Montréal, Québec, Canada.

Mende, M., Scott, M., Van Doorn, J., Grewal, D., & Shanks, I. (2019). Service robots rising. *Journal of Marketing Research, 56*(4), 535–556.

Millet, K., Buehler, F., Du, G., & Kokkoris, M. (2023). Defending humankind: Anthropocentric bias in the appreciation of AI art. *Computers in Human Behavior, 143*, 107707.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115(2)*, 502–517.

Mori, M. (1970). The uncanny valley. *Energy, 7*, 33–35.

Mori, M., MacDorman, K. F., Kageki, N. (2012). The uncanny valley [From the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98–100.

Moussawi, S., & Koufaris, M. (2019, January). *Perceived intelligence and perceived anthropomorphism of personal intelligent agents: Scale development and validation*. Paper presented at the meeting of Proceedings of the 52nd Annual Hawaii in International Conference on System Sciences, Hawaii, United State.

Nass, C., Steuer, J., & Tauber, E. R. (1994, April). *Computers are social actors*. Paper presented at the meeting of Conference on Human Factors in Computing Systems,Boston, Massachusetts, United State.

Natarajan, M., & Gombolay, M. (2020, March). *Effects of anthropomorphism and accountability on trust in human robot interaction*. Paper presented at the meeting of 2020 15th ACM/IEEE International Conference on Human–Robot Interaction (HRI), Cambridge, United Kingdom.

Nefdt, R. (2020). A puzzle concerning compositionality in machines. *Minds and Machines (Dordrecht), 30*(1), 47–75.

Newman, D., Fast, N., & Harmon, D. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes, 160*, 149–167.

Nicholson Price, W. (2018). Big data and black-box medical algorithms. *Science Translational Medicine, 10*(471). Article aao5333.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes.

*Psychological Review, 84*(3), 231–259.

Nørskov, S., Damholdt, M., Ulhøi, J., Jensen, M., Ess, C., & Seibt, J. (2020). Applicant fairness perceptions of a robot‑mediated job interview: A video vignette–based experimental survey. *Frontiers in Robotics and AI, 7*, 586263.

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2021, May). *Human–AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens*. Paper presented at the meeting of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Boston: Harvard University Press.

Prahl, A., & van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting, 36*(6), 691–702.

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141.

Schoeffer, J., Machowski, Y., & Kuehl, N. (2022). *Perceptions of fairness and trustworthiness based on explanations in human vs. automated decision-making*. Paper presented at the meeting of the Annual Hawaii International Conference on System Sciences, .Hawaii, United State.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47(5),* 609-612.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media, 64*(4), 541–565.

Shin, D., & Park, Y. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*, 277–284.

Smith, B., & Shum, H. (2018). *The future computed: Artificial intelligence and its role in society*. Independently Published By Microsoft.

Swart, J. (2021). Experiencing algorithms: How young people understand, feel about, and engage with algorithmic news selection on social media. *Social Media Society, 7*(2), 205630512110088.

Tomprou, M., & Lee, M. (2022). Employment relationships in algorithmic management: A psychological contract perspective. *Computers in Human Behavior, 126*, 106997.

Upadhye, C. (2018). How algorithms run Amazon`s warehouse. *Medium*. Retrieved December 27, 2018, from https://charviupadhye.medium.com/how-algorithms-run-amazons-warehouse-61e620ad27a7

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232.

Wen, Z., Zhang, L., Hou, J., & Liu, H. (2004). Testing and application of the mediating effects. *Acta Psychological Sinica, 36*(5), 614–620.

[温忠麟, 张雷, 侯杰泰, 刘红云. (2004). 中介效应检验程序及其应用. *心理学报, 36*(5), 614–620.]

Wesche, J. S., & Sonderegger, A. (2019). When computers take the lead: The automation of leadership. *Computers in Human Behavior, 101*, 197–209

Wien, A., & Peluso, A. (2021). Influence of human versus AI recommenders: The roles of product type and cognitive processes. *Journal of Business Research, 137*, 13–27.

Wu, M., Wang, N., & Yuen, K. (2023). Deep versus superficial anthropomorphism: Exploring their effects on human trust in shared autonomous vehicles. *Computers in Human Behavior, 141*, 107614.

Xu, L., & Yu, F. (2020). Factors that influence robot acceptance. *Chinese Science Bulletin, 65*(6), 496–510.

[许丽颖, 喻丰. (2020). 机器人接受度的影响因素. *科学通报, 65*(6), 496–510.]

Xu, L., Yu, F., & Peng, K. (2022). Algorithmic discrimination causes less desire for moral punishment that human discrimination. *Acta Psychological Sinica, 54*(9), 1076–1092.

[许丽颖, 喻丰, 彭凯平. (2022). 算法歧视比人类歧视引起更少道德惩罚欲. *心理学报, 54*(9), 1076–1092.]

Xu, L., Yu, F., Wu, J., Han, T., & Zhao., L. (2017). Anthropomorphism: Antecedents and consequences. *Advances in Psychological Science, 25*(11), 1942–1954.

[许丽颖, 喻丰, 邬家骅, 韩婷婷, 赵靓. (2017). 拟人化: 从"它"到"他". *心理科学进展, 25*(11), 1942–1954.]

Yam, K., Bigman, Y., Tang, P., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2021). Robots at work: People prefer–and forgive–service robots with perceived feelings. *Journal of Applied Psychology, 106*(10), 1557–1572.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making, 32*(4), 403–414.

Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human–Robot Interaction, 5*(2), 29−47.

Yu, H., Miao, C., Chen, Y., Fauvel, S., Li, X., & Lesser, V. (2017). Algorithmic management for improving

collective productivity in crowdsourcing. *Scientific Reports, 7*(1), 12541.

Yuan, L., & Dennis, A. (2019). Acting like humans? Anthropomorphism and consumer's willingness to pay in electronic commerce. *Journal of Management Information Systems, 36*(2), 450–477.

Zajonc, R. B. (1968). Attitude effects of mere exposure. *Journal of Personality and Social Psychology, 9*(2), 1–27.

Zhang, Y., Xu, L., Yu, F., Ding, X., Wu, J., & Zhao, L. (2022). A three-dimensional motivation model of algorithm aversion. *Advances in Psychological Science, 30*(5), 1093–1105.

[张语嫣, 许丽颖, 喻丰, 丁晓军, 邬家骅, 赵靓. (2022). 算法拒绝的三维动机理论. *心理科学进展, 30*(5), 1093-1105.]

Zhao, T. (2018). Comprehensive considerations for the "revolution" of artificial intelligence: An ethical and ontological analysis. *Philosophical Trends,* (4), 5–12.

[赵汀阳. (2018). 人工智能"革命"的"近忧"和"远虑"——一种伦理学和存在论的分析. *哲学动态,* (4), 5–12.]

Zhao, T. (2019). How could AI develop its self-consciousness? *Journal of Dialectics of Nature, 41*(1), 1–8.

[赵汀阳. (2019). 人工智能的自我意识何以可能？ *自然辩证法通讯, 41*(1), 1–8.]

Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human–Computer Studies, 100*, 48−54.

# Perceived opacity leads to algorithm aversion in the workplace

ZHAO Yijun, XU Liying, YU Feng, JIN Wanglong

(*Department of Psychology, Wuhan University, Wuhan 430072, China*)

**Abstract**

With algorithms standing out and influencing every aspect of human society, people's attitudes toward algorithmic invasion have become a vital topic to be discussed. Recently, algorithms as alternatives and enhancements to human decision-making have become ubiquitously applied in the workplace. Despite algorithms offering numerous advantages, such as vast data storage and anti-interference performance, previous research has found that people tend to reject algorithmic agents across different applications. Especially in the realm of human resources, the increasing utilization of algorithms forces us to focus on users' attitudes. Thus, the present study aimed to explore public attitudes toward algorithmic decision-making and probe the underlying mechanism and potential boundary conditions behind the possible difference.

To verify our research hypotheses, four experiments ($N$ = 1211) were conducted, which involved various kinds of human resource decisions in the daily workplace, including resume screening, recruitment and hiring, allocation of bonuses, and performance assessment. Experiment 1 used a single-factor, two-level, between-subjects design. 303 participants were randomly assigned to two conditions (agent of decision-making: human versus algorithm) and measured their permissibility, liking, and willingness to utilize the agent. Experiment 1 was designed to be consistent with experiment 2. The only difference was an additional measurement of perceived transparency to test the mediating role. Experiment 3 aimed to establish a causal chain between the mediator and dependent variables by manipulating the perceived transparency of the algorithm. In experiment 4, a single-factor three-level between-subjects design (non-anthropomorphism algorithm versus anthropomorphism algorithm versus human) was utilized to explore the boundary condition of this effect.

As anticipated, the present research revealed a pervasive algorithmic aversion across diverse organizational settings. Specifically, we conceptualized algorithm aversion as a tripartite framework encompassing cognitive, affective, and behavioral dimensions. We found that

compared with human managers, participants demonstrated significantly lower permissibility (Experiments: 1, 2, and 4), liking (Experiments: 1, 2, and 4), and willingness to utilize (Experiment 2) algorithmic management. And the robustness of this result was demonstrated by the diversity of our scenarios and samples. Additionally, this research discovered perceived transparency as an interpretation mechanism explaining participants' psychological reactions to different decision-making agents. That is to say, participants were opposed to algorithmic management because they thought its decision processes were more incomprehensible and inaccessible than humans (noted in Experiment 2). Addressing this "black box" phenomenon, experiment 3 showed that providing more information and principles about algorithmic management positively influenced participants' attitudes. Crucially, the result also demonstrated the moderating effect of anthropomorphism. The result showed that participants exhibited greater permissibility and liking for the algorithm with human-like characteristics, such as a human-like name and communication style, over more than a mechanized form of the algorithm. This observation underlined the potential of anthropomorphism to ameliorate resistance to algorithmic management.

These results bridge the gap between algorithmic aversion and decision transparency from the social-psychological perspective. Firstly, the present research establishes a three-dimensional (cognitive, affective, and behavioral) dual-perspective (employee and employer) model to elucidate the negative responses toward algorithmic management. Secondly, it reveals that perceived opacity acts as an obstacle to embracing algorithmic decision-making. This finding lays the theoretical foundation of Explainable Artificial Intelligence (XAI) which is conceptualized as a "glass box". Ultimately, the study highlights the moderating effect of anthropomorphism on algorithmic aversion. This suggests that anthropomorphizing algorithms could be a feasible approach to facilitate the integration of intelligent management systems.

**Key words** algorithm aversion, transparency, anthropomorphism, workplace